Visualizing the Impact of Chemical Substructures on Compound Activity for Improving the Drug Discovery Process

Jochem Nelen[1], Miguel Carmena-Bargueño[1], Carlos Martínez-Cortés[1], Alejandro Rodríguez-Martínez[1], Antonio Jesús Banegas-Luna[1], Alfonso Pérez-Garrido[1], Jose Manuel Villalgordo-Soto[2] and Horacio Pérez-Sánchez[1]

[1] Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), HiTech Innovation Hub, Universidad Católica de Murcia (UCAM), Spain
[2] Eurofins Villapharma Research, Murcia, Spain

UCAM
UNIVERSIDAD CATÓLICA DE MURCIA

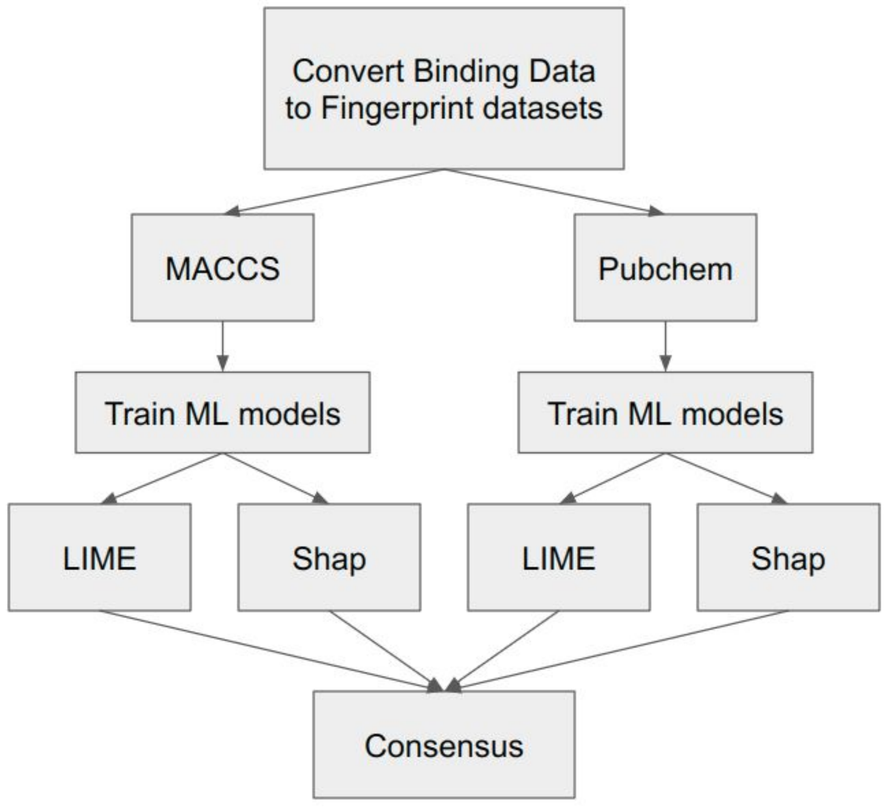
Abstract

- Machine learning (ML) can be a powerful tool in drug discovery but often operates as a “black box”
- Interpretation techniques exist which can provide insight in *why* the ML model predicts certain things
- This concept was applied to a drug-discovery context using Sibila, where interpretation of the ML models can indicate which substructures can be beneficial for compound activity

Interpretation Methods

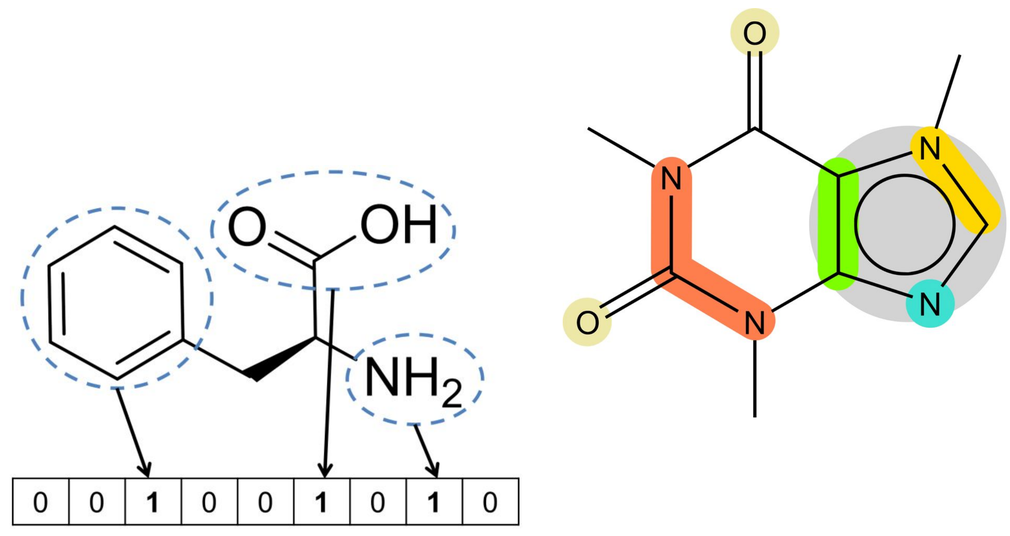
- LIME
- Shap
- …
- Different ways of identifying important features → combine to get a better picture
- Train and interpret models using Sibila

Workflow



Fingerprints

- MACCS keys (166 features)
- Pubchem Fingerprint (881 features)
- If feature substructure is present:
  → feature bit is set to 1, else to 0



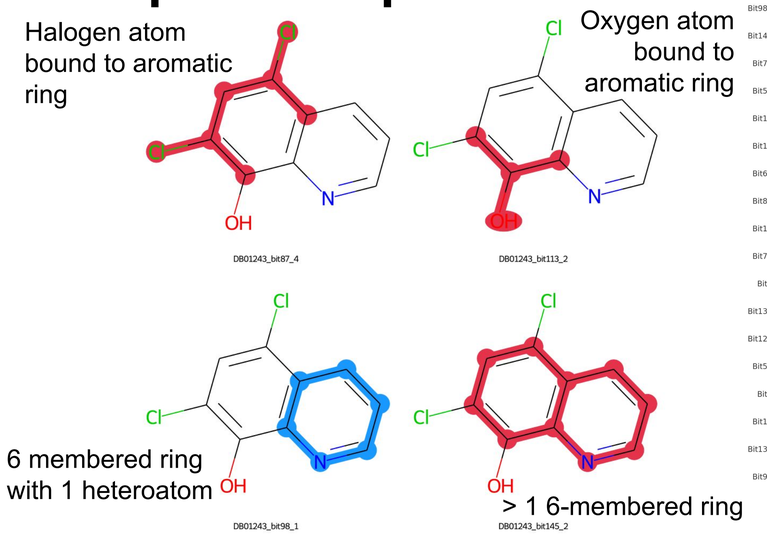
Output example from ANN

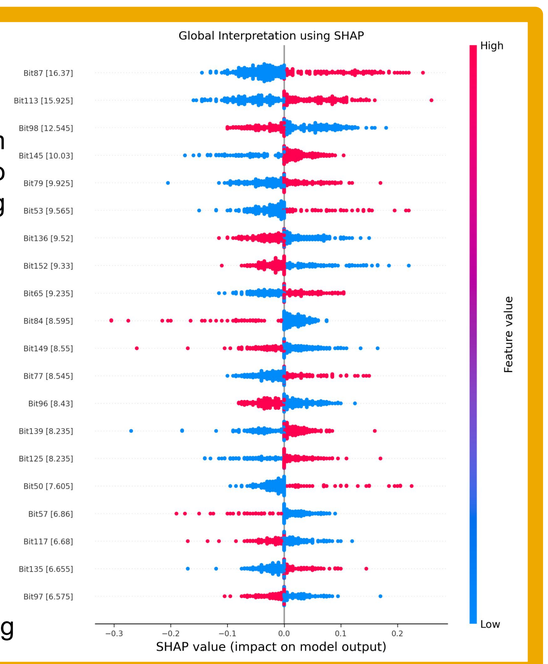
Halogen atom bound to aromatic ring

Oxygen atom bound to aromatic ring

6 membered ring with 1 heteroatom

> 1 6-membered ring





Conclusions

- ML-based Virtual Screening is powerful in drug discovery
- Interpretation techniques combined with substructure fingerprints aid in model interpretation
- Consensus of interpretation results from different fingerprints and models provides robust insights
- Sibila facilitates training and interpretation of models

Acknowledgments and Funding

J.N. holds a PhD fellowship at UCAM funded by Cátedra Eurofins VillaPharma.

Villapharma research
sigma2
PSNC
CENTRO EXTREMEÑO DE TECNOLOGÍAS AVANZADAS CETA Ciemat
NLHPC National Laboratory for High Performance Computing Chile

References

1. Sibila: https://github.com/bio-hpc/sibila
2. LIME: https://doi.org/10.1145/2939672.2939778
3. Shap: https://dl.acm.org/doi/10.5555/3295222.3295230
4. MACCS key: https://doi.org/10.1021/ci010132r
5. Pubchem FP: https://doi.org/10.1186/s13321-017-0195-1

Contact: jnelen@ucam.edu

BIO-HPC Structural Bioinformatics and High Performance Computing Research Group

UCAM | UNIVERSIDAD CATÓLICA DE MURCIA